基于信号检测论的认知诊断评估：构建与应用[1]

郭磊[1,2] 秦海江[1,3]

（[1]西南大学心理学部，重庆 400715）

（[2]中国基础教育质量监测协同创新中心西南大学分中心，重庆 400715）

（[3]贵阳市第三十七中学，贵阳 550003）

摘要 作答选择题可被看作从噪音中提取信号的过程，研究提出了一种基于信号检测论的认知诊断模型(SDT-CDM)。新模型的优势在于：(1)无需对选项进行属性层面的编码。(2)能获得传统诊断模型无法提供的题目区分度和难度参数。(3)可以直接表达每个选项之间的合理性差异，对题目性能刻画更加细微全面。两个模拟研究结果表明：(1)EM 算法可以实现对新模型的参数估计过程，便捷有效。(2)SDT-CDM 具备良好性能，分类准确性和参数估计精度较高以外，还能提供选项层面的估计信息，用于题目质量诊断与修订。(3)属性数量、题目质量与样本量等因素会影响 SDT-CDM 的表现。(4)与称名诊断模型 NRDM 相比，SDT-CDM 在所有实验条件下对被试的分类准确性更高。实证研究表明：SDT-CDM 比 NRDM 具有更好的模型数据拟合结果，其分类准确性和一致性更高，尤其当属性考察次数较少时具有很强的稳定性，难度和区分度参数与 IRT 模型估计结果的相关性也更高，值得推广。

关键词 信号检测论，认知诊断，选择题，EM 算法

1 引言

自 Kelly(1916)第一次提出选择题(Multiple-Choice, MC)测验形式，因其客观、有效、便捷等特点而广受欢迎，直至当下仍是测验主流题型之一，并广泛应用在 TIMSS、PISA、NAEP 和 TOEFL 等标准化测验。MC 题型具有诸多优势：不受主观误差影响、提高测验信度、易于批阅且计分快速、满足内容平衡需求等(郭磊, 周文杰, 2021)。通常，MC 作答数据被当作 0-1 计分形式(即答对或答错)处理，但这样会造成干扰项信息的损失。为了充分挖掘干扰项的诊断信息，提高个体知识状态的分类精度，研究者提出了许多方法，如 MC-DINA 模型(Multiple-Choice DINA; de la Torre, 2009)及其拓展的结构化 MC-DINA 模型(Ozaki, 2015)，包含干扰项信息的 SICM 模型(Scaling Individuals and Classifying Misconceptions Model;

[1] 基金项目：国家自然科学基金青年项目（31900793）；中央高校基本科研业务费专项资金（SWU2109222）；西南大学 2035 先导计划项目（SWUPilotPlan006）

通讯作者：郭磊，E-mail: happygl1229@swu.edu.cn

收稿日期：2023 年 4 月 21 日

Bradshaw & Templin, 2014)和 GDCM-MC 模型(Generalized Diagnostic Classification Models for Multiple Choice Option-Based Scoring; DiBello et al., 2015)，以及基于选项层面的非参数认知诊断方法(郭磊 等, 2021; Wang et al., 2023)。这些方法的目标是在知识状态空间中对被试进行分类，从而知晓其学科知识或认知属性的掌握情况，这种评估方式也被称作认知诊断评估。但上述 MC 处理方法有个前提条件，即要求对干扰项进行编码，然后才能表征出区别于正确选项所表征的潜在类别。虽然前期的研究要求干扰项的编码需要是正确选项编码的子集、不同干扰项之间也要有包含关系(郭磊 等, 2013)，但最近的研究已突破了该限制，即干扰项的编码无需嵌套于正确选项编码中(Wang et al., 2023)，进一步推动了该领域研究。

实际上，MC 测验也可以被视作一种信号检测任务，被试需从一系列的噪音(所有选项)背景中选择出信号，即做出正确反应。被试作答过程中存在两种可能性，要么“会答/知道(Know)”，要么“不会答/不知道(do not Know)”。从信号检测论(signal detection theory, SDT)的视角出发，被试作答行为可包含两个阶段：①感知阶段：被试在理解题意后对每个选项产生不同程度的合理性[2](plausibility)判断，可用合理性参数表达，每个选项的合理性参数均服从一定分布。②决策阶段：被试在权衡每个选项的合理性后，会做出选择最合理选项的决策。基于该理念，DeCarlo(2021)将 SDT 与项目反应理论(IRT)结合用于 MC 题目分析，通过 SDT 模型可获得被试在选择各选项时的相对合理性参数、以及题目的区分度和难度参数信息。研究表明，SDT 模型估计得到的难度参数与两参数、三参数项目反应模型基本一致，但区分度参数仅与两参数模型相关较高，与三参数模型相关低至 0.04。此外，SDT 还可以提供更丰富的信息，如被试对每个选项尤其是干扰项的合理性倾向，以及被试在每个选项上感知到的合理性差异(即选项差异)。因此，SDT 对题目的解析更细微，可以从选项层面知晓题目的整体情况，其价值在于：①若某道题目偏简单，为了增大该题目难度，可以通过估计得到的选项合理性参数进行选项层面的针对性调整，起到修订题目的作用。②诊断题目是否有问题。当被试“会答”该题目时，选择干扰项的倾向性仍比选择正确选项的倾向性更大，则预示着该题目的质量出现了问题。以上优势是两参数和三参数模型无法做到的。此外，SDT 对 MC 题目的分析要比称名反应模型(Nominal Response Model, NRM; Bock, 1972)更加简洁易于解释。尽管 NRM 也可分析基于选项的数据，但它引入了多个区分度参数，使得参数估计和结果解释都变得复杂。若进一步想在 NRM 中表征猜测行为的话，又需要引入更多的猜测参数，这会导致模型参数增多并且难以估计(Thissen & Steinberg, 1997)，但 SDT 模型无需增加额外

[2] 合理性可理解为基于个人知识、经验等因素认为该选项是正确的/合理的倾向性。

参数便可对猜测行为进行表征，更加简约。并且根据 DeCarlo(2021)的实证研究[3]表明，SDT 模型比 NRM 有更好的模型拟合结果。

尽管在认知诊断评估中，Templin 等(2008)将 NRM 拓展为称名反应诊断模型(Nominal Response Diagnostic Model, NRDM)，使之能够分析认知诊断的数据。随后，Ma 和 de la Torre(2016)提出了顺序 G-DINA(sequential G-DINA)的模型框架，将 NRDM 包含在内，可实现对顺序(ordered)和称名数据的处理。但这些模型均是基于最初 NRM 思想的拓展，也保留了 NRM 存在的问题，如题目参数过多等问题：每道题目的每个选项都要估计截距项、主效应项及其交互作用项。因此，基于 SDT 视角分析选项层面的诊断数据，并探讨其适用价值具有重要意义。SDT 用于认知诊断评估有以下优势：①无需对 MC 题目的选项进行编码，节省大量人力物力。②在保证提供选项水平分析结果的前提下，还可以使用更加精简的模型表达方式来达到比 NRDM 模型更好的解释意义，参数更容易估计。③由于模型更加简洁，模型和数据的拟合可能会进一步提升。④能够提供传统诊断模型无法提供的难度和区分度[4]参数。

综上所述，信号检测论视角的 MC 题型认知诊断评估将具备诸多优势，因此本文拟探讨基于信号检测论的 MC 题型认知诊断评估方法与技术，构建 SDT-CDM 模型并推导其参数估计方法，并在模拟和实证测验中检验新模型的性能和有效性。本文结构如下：首先介绍 SDT 模型的逻辑背景，其次阐述 SDT 诊断模型(记作 SDT-CDM)的构建过程和参数估计方法，之后通过模拟和实证研究探讨 SDT-CDM 的性能，最后对结果进行讨论与展望。

2 SDT 模型简介

被试在作答 MC 题目时，首先会对每个选项产生不同程度的感知，进而将这种感知转换成认为该选项是正确答案的合理性倾向。为了用模型表达出该加工过程，可认为被试对每个选项的合理性倾向均服从一个概率分布，如图 1 所示。

[3] 600 名被试参与的 32 道题目的学术评估测试(Scholastic Assessment Test, SAT)，每道题目有 5 个选项。
[4] 传统诊断模型没有难度参数的具体表达，而区分度是通过估计得到参数后计算才能得到。

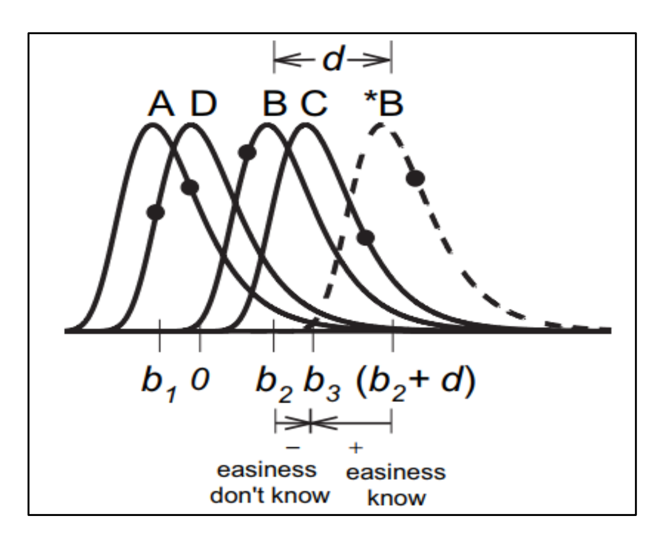


图 1 SDT 模型的反应示意图(取自 DeCarlo, 2021; P3, Figure 1)

图 1 呈现了 SDT 的反应过程，假设某四选一的 MC 题目，选项分别是 A、B、C 和 D，正确答案为 B。一方面，若被试不会作答该题目，他就会凭借自己感知到各个选项的合理性作答，感知越合理的选项其分布越靠右，如图 1 实线分布所示。当感知之后便是决策过程：被试会选择感知到合理性最强的那个选项，在该例子中被试感知选项合理性的大小依次为：C>B>D>A，即被试最有可能选择 C 选项，其位置最靠右端。为了实现模型参数估计，需要固定某个选项的合理性参数作为参照组，通常将最后一个选项 D 的参数固定为 0。因此，各选项合理性的相对差异大小可由各选项离开“0”的距离进行表示，该例子中 A、B 和 C 选项的合理性参数被标记为b_1、b_2和b_3，它们可类比为多级计分模型中的阈值参数。另一方面，若被试“会答”该题目，那么被试对正确选项 B 感知到的合理性就最强，正确选项 B 的分布就会漂移至最右端，即图中虚线分布位置，由于此时 B 处于最右端，因此被试将做出“选择正确答案 B”的反应。

如图 1 所示，B 和*B 分布之间的距离d可作为被试“会答”和“不会答”时选择正确选项的差异，即题目区分度参数d，该参数与 IRT 中区分度a参数作用相似。显然，d越大，题目的区分度越高。若d为负值，表明题目存在问题：“不会答”该题的被试反而比“会答”该题目的被试更容易答对题目，可以考虑修改或删除该题。此外，DeCarlo(2021)根据被试“不会答”与“会答”题目的情况，定义了题目的两个易度参数[5]，即e_{DK} (easiness don’t-know) 与e_K (easiness know)，两者含义均为被试感知到的正确选项的合理性与剩余最高的合理性之间的差值。具体地，如图 1 示例：①若被试“不会答”题目，其感知到的 A、B、C 与 D 选项

[5] SDT 模型中的两类易度参数与项目反应模型中的难度 b 参数为反向理解，即易度参数取值越大，表明题目越简单。

的合理性分别为b_1、b_2、b_3和b_4，此时的易度参数为e_{DK}，并且有$e_{DK} = b_2 - b_3$；②若被试“会答”题目，其感知到的四个选项的合理性分别为b_1、$b_2 + d$、b_3和b_4，此时的易度参数为e_K，并且有$e_K = b_2 + d - b_3 = e_{DK} + d$。

SDT 模型中的区分度 d 可用于衡量题目质量，并且有$d = e_K - e_{DK}$。对于e_{DK}而言，当e_{DK}为负且越小时，表明“不会答”题目的被试选择正确选项的概率越小，“不会答”的被试更可能选择干扰项，符合测验逻辑；而当e_{DK}为正且越大时则违反测验逻辑。而对于e_K而言，有与e_{DK}相反的含义：当e_K为正且越大时，表明“会答”题目的被试选择正确选项的概率越大，“会答”的被试更可能选择正确选项，符合测验逻辑；而当e_K为负且越小时则违反测验逻辑。因此，当e_{DK}为正且越大或e_K为负且越小时，题目质量都存在问题，可以考虑修改或者删除该题目。

通过上述对d、e_{DK}、e_K等参数作用的理解，可以感受到 SDT 模型在评价题目质量与指导题目修改方面的优势。在实际测验中，MC 题目质量通常无法保证，即使大型测验也会出现猜测概率较高的情况，如 DeCarlo(2021)分析的 32 道 SAT12 题目中，就有 17 道题目的e_{DK}为正、2 道题目的e_K为负的情况出现。通过 SDT 模型可以简单高效地筛选出有问题的题目，并且能够指导题目的修改，非常有价值。

基于上述理论基础，SDT 模型本质上是一个混合模型，如公式(1)所示(详细推导请参见 DeCarlo(2021))：

$$P_{jm}\left(b_{jm}, d_j, X_{jm}, \lambda_i\right) = \lambda_i \frac{e^{b_{jm}+d_j X_{jm}}}{\sum_{h=1}^{M} e^{b_{jh}+d_j X_{jh}}} + (1-\lambda_i)\frac{e^{b_{jm}}}{\sum_{h=1}^{M} e^{b_{jh}}} \quad (1)$$

其中，P_{jm}表示被试在第 j 题上选择第m个选项($m = 1, \dots, M$)的概率，M为 MC 题目选项的总个数。λ_i是一个混合参数，用以表示被试 i 会作答题目的概率，取值在 0-1 之间。b_{jm}/b_{jh}为题目在第m/h个选项上的合理性参数。d_j为题目j的区分度。X_{jm}/X_{jh}为示性函数，表示选项m/h是否为正确答案，若是则取 1，否则取 0。模型前半部分表示被试会作答题目时，选择第m个选项占所有选择可能性的百分比，后半部分表示被试不会作答题目时的情况。

3 SDT-CDM 的构建及参数估计

为了将 SDT 用于认知诊断评估，构建出基于信号检测论的认知诊断模型 SDT-CDM，需要满足：①在 SDT 模型中表征出被试的知识状态用以进行分类诊断。②被试知识状态和题目$\boldsymbol{q}$向量之间的相互作用需要反映在模型中，并且不同的知识状态对会作答题目的影响应当不同，以实现模型对不同知识状态被试的识别。③模型要可识别，并且可以通过常用的估

计算法，如EM或MCMC算法实现模型的参数估计。基于以上三点，本研究提出了SDT-CDM，如公式(2)所示。

$$P_{jm}(b_{jm},d_j,X_{jm},\boldsymbol{\alpha}_l)=\lambda_{lj}\frac{e^{b_{jm}+d_jX_{jm}}}{\sum_{h=1}^{M}e^{b_{jh}+d_jX_{jh}}}+\left(1-\lambda_{lj}\right)\frac{e^{b_{jm}}}{\sum_{h=1}^{M}e^{b_{jh}}} \quad (2)$$

其中，$\boldsymbol{\alpha}_l$表示知识状态为第 l 种类别(l = 1,2,…,2^K)，K为属性个数。λ_{lj}表示知识状态为$\boldsymbol{\alpha}_l$的被试会作答题目 j 的概率，与公式(1)不同，SDT-CDM 的优势可以刻画被试与不同$\boldsymbol{q}$向量类型的题目之间的交互作用，同时放松了传统 SDT 模型仅能反映被试总体水平(即λ_i)而不是反映被试与具体题目之间的交互作用信息的强假设，使模型更灵活。其余符号同公式(1)。混合参数λ_{lj}的计算如下所示：

$$\lambda_{lj}=\frac{\sum_{k=1}^{K_j^*}\delta_{jk}\alpha_{lk}+\sum_{k=1}^{K_j^*-1}\sum_{k'=k+1}^{K_j^*}\delta_{jk'k}\alpha_{lk}\alpha_{lk'}+\cdots+\delta_{j12\ldots K_j^*}\prod_{k=1}^{K_j^*}\alpha_{lk}}{\sum_{k=1}^{K_j^*}\delta_{jk}q_{jk}+\sum_{k=1}^{K_j^*-1}\sum_{k'=k+1}^{K_j^*}\delta_{jk'k}q_{jk}q_{jk'}+\cdots+\delta_{j12\ldots K_j^*}\prod_{k=1}^{K_j^*}q_{jk}} \quad (3)$$

尽管λ_{lj}的分子部分构造与 G-DINA 模型(de la Torre，2011)类似，但参数的含义完全不同。首先，λ_{lj}的计算中不存在截距项，意味着当被试未掌握题目考察的任何属性时，其值为 0，即表示不会作答，与 SDT 模型理念保持一致。此时，虽然δ_{jk}也可被看作题目j的第k个属性的主效应，但其含义为：被试掌握了题目j所考察的第k个属性时，对于“会作答”该题目的概率的贡献程度，而非对正确作答概率的贡献程度，这是 SDT-CDM 与 G-DINA 的本质区别。$\delta_{jk'k}$为题目的二阶交互项，$\delta_{j12\ldots K}$为最高阶交互项，其含义与主效应含义类似。Kj^*为题目 j 考察到的属性个数。λ_{lj}的分母表示题目 j 考察的全部属性的效应之和，分子表示被试掌握了其中的部分或全部属性的效应之和，因此，被试掌握所考察的属性越多，那么其“会作答”的概率也就越高，不同类型的知识状态，对于相同题目会作答的概率是不同的，这也是 SDT-CDM 优于 SDT 模型的其中一点。

不难看出，若被试未掌握题目 j 考察的任何属性时，有$\lambda_{\boldsymbol{0}j}=0$，若被试掌握了所有考察属性时，有$\lambda_{\boldsymbol{1}j}$=1。因此，有$\lambda_{lj}=\frac{\lambda_{lj}}{1}=\frac{\lambda_{lj}}{\lambda_{\boldsymbol{1}j}}$，公式(3)可改写为：

$$\lambda_{lj}=\sum_{k=1}^{K_j^*}\delta_{jk}\alpha_{lk}+\sum_{k=1}^{K_j^*-1}\sum_{k'=k+1}^{K_j^*}\delta_{jk'k}\alpha_{lk}\alpha_{lk'}+\cdots+\delta_{j12\ldots K_j^*}\prod_{k=1}^{K_j^*}\alpha_{lk} \quad (4)$$

SDT-CDM 的模型参数估计可用 MMLE/EM 算法实现，算法推导过程及其标准误计算请参见网络版附录。

4 模拟研究 1

4.1 研究目的

采用蒙特卡洛模拟方式探讨 SDT-CDM 在不同实验条件下对被试的分类准确性和参数估计精度。

4.2 实验设计

本研究为 5 因素完全交叉设计，5 个自变量分别为属性个数(K = 3, 5)、题目长度(J = 20, 40)、题目质量(高质量，低质量)、样本量(N = 1000, 2000)、属性分布(高阶分布，多元正态分布)。所有实验条件均重复 200 次以减少随机误差。

4.2.1 题目的模拟

Q 矩阵的生成方式为：在保证有两个单位矩阵的情况下，其余题目的$\boldsymbol{q}$向量从所有可能的$\boldsymbol{q}$向量中随机抽取，以实现被试知识状态的可识别(Xu, 2017; Fang et al., 2019)与 Q 矩阵的随机模拟。题目质量由于没有前人研究作为参照，因此参考之前认知诊断相关研究中的范围进行设置(郭磊 等, 2016)，具体为：高质量题目参数中 e_{DK} 从$U[-2.5, -1]$随机抽取且 e_K 从$U[2.5, 3.5]$中随机抽取，由于当 e_{DK} = -2.5 且 e_K = 3.5 时有$(1-P_{\mathbf{1}})$和$P_{\mathbf{0}} \cong 0.05$，当 e_{DK} = -1 且 e_K = 2.5 时有$(1-P_{\mathbf{1}})$和$P_{\mathbf{0}} \cong 0.15$，此时与认知诊断中$(1-P_{\mathbf{1}})$和$P_{\mathbf{0}}$ 从$U[0.05, 0.15]$中随机抽取等价；低质量题目参数中 e_{DK} 从$U[-1, -0.5]$随机抽取且 e_K 从$U[1.8, 2.5]$中随机抽取，由于当 e_{DK} = -1 且 e_K = 2.5 时有$(1-P_{\mathbf{1}})$和$P_{\mathbf{0}} \cong 0.15$，当 e_{DK} = -0.5 且 e_K = 1.8 时有$(1-P_{\mathbf{1}})$和$P_{\mathbf{0}} \cong 0.25$，此时与$(1-P_{\mathbf{1}})$和$P_{\mathbf{0}}$ 从 $U[0.15, 0.25]$随机抽取等价。为了最大程度实现模拟数据的随机性与结论的可推广性，题目的合理性参数b_{jm}与属性效应δs 均不做严格约束。由于合理性参数b_{jm}仅通过相对大小来影响选择某选项的概率(如图 1 所示)，因此可从标准正态分布中随机抽取，以实现选项之间合理性倾向的随机大小关系。属性效应δs 满足“掌握属性越多的被试其‘会答’题目的概率越高”这一假设即可。此外，本研究固定 MC 题目的选项数量为 4 个，与现实中大多数 MC 题型的选项数量一致。

4.2.2 被试的模拟

被试的知识状态采取高阶和多元正态分布生成。其中，高阶分布参考 Ma 等(2016)的设置，具体如下：

$$P_{ik} = \frac{\exp[1.7 \times (\theta_i - \delta_k)]}{1 + \exp[1.7 \times (\theta_i - \delta_k)]} \qquad (5)$$

其中，θ_i为被试的能力大小，从标准正态分布中抽取。δ_k为属性k被掌握的难度，从-1 到 1 之间按照属性数量等距选取，如 3 属性时三个属性的难度分别为$\delta_1 = -1, \delta_2 = 0, \delta_3 = 1$。

多元正态分布参考 Chiu(2013)的设置，定义一个 K 维向量$\boldsymbol{\theta}_i = (\theta_{i1}, \theta_{i2} \dots, \theta_{iK})$作为被试 i 在每个属性上的连续能力值，$\boldsymbol{\theta}_i$从多元正态分布$MVN(\mathbf{0}, \boldsymbol{\Sigma})$中随机抽取，协方差矩阵$\boldsymbol{\Sigma}$的非对角线元素用于描述属性间的相关，设置为 0.5。被试知识状态真值可用下式生成：

$$\alpha_{ik} = \begin{cases} 1 \ , if \ \ \theta_{ik} > \varphi^{-1}(\dfrac{k}{1+K}) \\ 0 \ , \qquad otherwise \end{cases} \qquad (6)$$

4.3 评价指标

参数估计精度的评价指标主要采用平均偏差 Bias、均方误差根(root mean squared error, RMSE)，计算见公式(7)和公式(8)。

$$Bias = \frac{\sum_{r=1}^{R}(\omega - \widehat{\omega}_r)}{R} \qquad （7）$$

$$RMSE = \sqrt{\frac{\sum_{r=1}^{R}(\omega - \widehat{\omega}_r)^2}{R}} \qquad （8）$$

其中，ω表示参数“真值”，$\widehat{\omega}_r$表示参数估计值，R表示总循环次数，r表示当前循环次数。Bias 越接近于零表明参数估计的偏差性越小，RMSE 越小说明参数估计的准确性越好。

被试属性掌握情况的估计精度评价指标采用平均属性判准率(average attribute correct classification rate, AACCR)和模式判准率(pattern correct classification rate, PCCR)，计算公式如下：

$$AACCR = \frac{\sum_{k}^{K} ACCR_k}{K} \qquad (9)$$

$$PCCR = \frac{\sum_{r=1}^{R}\sum_{i=1}^{N} pm_{ir}}{R \times N} \qquad (10)$$

其中，$ACCR_k = \frac{\sum_{r=1}^{R}\sum_{i=1}^{N} \alpha m_{ikr}}{R \times N}$ （11）

其中，$\alpha m_{ikr} = 1$表示第r次循环中对被试 i 的第 k 个属性判断正确，$pm_{ir} = 1$表示第r次循环中被试 i 的知识状态判断正确。

4.4 研究结果

图 2 和图 3 呈现了不同自变量水平下 SDT-CDM 的参数估计 Bias 和 RMSE 的总体结果。由于每道题目的合理性参数、属性主效应和属性交互效应的参数不止一个，考虑呈现的简洁性和篇幅，结果用均值表示。整体来看，各参数的估计精度均较高，如：合理性参数 Bias 范围为-0.003 至 0.007，均值为 0.002；RMSE 范围为 0.119 至 0.261，均值为 0.173。区分度参数 Bias 范围为-0.054 至-0.001，均值为-0.022；RMSE 范围为 0.145 至 0.385，均值为 0.253。

易度参数 e_K 的 Bias 范围为-0.014 至 0.075，均值为 0.027；RMSE 范围为 0.181 至 0.334，均值为 0.260。其余参数不再赘述。

不同自变量对参数估计精度的影响不同。首先，属性分布为高阶分布的精度要稍优于多元正态分布的精度，如高阶分布下的 b、d、e_{DK}、e_K、δ-M 和 δ-I 参数的 Bias(RMSE)均值分别为 0.002(0.160)、-0.022(0.234)、0.046(0.245)、0.025(0.248)、0.001(0.078)和-0.001(0.154)，多元正态分布下的对应参数的 Bias(RMSE)均值分别为 0.002(0.187)、-0.022(0.271)、0.051(0.267)、0.029(0.271)、0.008(0.126)和-0.009(0.236)。其次，属性个数越多，精度会略有下降，如由 K=3 变为 K=5 时，尽管所有参数的 Bias 均值由 0.009 变为 0.010，但 RMSE 的均值由 0.189 增大至 0.224，增幅为 18.5%。然而，题目数量对参数估计精度的影响较小。当 J=20 增加至 40 题时，所有参数的 Bias 均值由 0.008 变为 0.010，RMSE 的均值由 0.203 变为 0.210，相差无几。再次，题目质量对精度的影响较大，当题目质量由高变低时，所有参数的 Bias 均值由 0.000 变为 0.019，RMSE 的均值由 0.192 变为 0.221，增幅为 15.1%。最后，样本量的影响最大，当人数由 2000 降低至 1000 时，所有参数的 Bias 均值由 0.007 变为 0.010，RMSE 的均值由 0.179 变为 0.234，增幅高达 30.7%。

图 4 呈现了 SDT-CDM 的 AACCR 和 PCCR 判准率结果。整体而言，新模型能够较为准确的对被试进行分类，其分类精度同样会受不同自变量的影响。在本文关注的 5 个因素中，对分类精度影响最大的是题目质量。当题目质量较低时，AACCR 的范围为 0.902 至 0.988，均值为 0.951，PCCR 的范围为 0.609 至 0.964，均值为 0.816；当题目质量提升后，AACCR 的范围为 0.973 至 1.000，均值为 0.990，PCCR 的范围为 0.876 至 0.999，均值为 0.957，增幅为 17.4%。其次是属性个数对精度的影响，当 K=3 时，AACCR 的范围为 0.950 至 1.000，均值为 0.983，PCCR 的范围为 0.858 至 0.999，均值为 0.951；当 K=5 时，AACCR 的下降幅度为 2.5%，而 PCCR 的下降幅度为 15.7%。第三位的影响因素为题目数量，题量越多，对被试获得的信息就越多，因此对其分类精度也会提升。如 J=20 时，平均的 AACCR 和 PCCR 分别为 0.958 和 0.841，当 J=40 时，平均的 AACCR 和 PCCR 分别提升至 0.984 和 0.932，增幅分别为 2.7%和 10.8%。而其余两个变量：属性分布和样本量对分类精度的影响不大。如高阶分布时的平均 AACCR 和 PCCR 分别为 0.969 和 0.882，多元正态分布时的平均 AACCR 和 PCCR 分别为 0.972 和 0.891；人数为 1000 人时的平均 AACCR 和 PCCR 分别为 0.970 和 0.883，当人数增长至 2000 时，平均 AACCR 和 PCCR 分别为 0.972 和 0.890，相差无几。

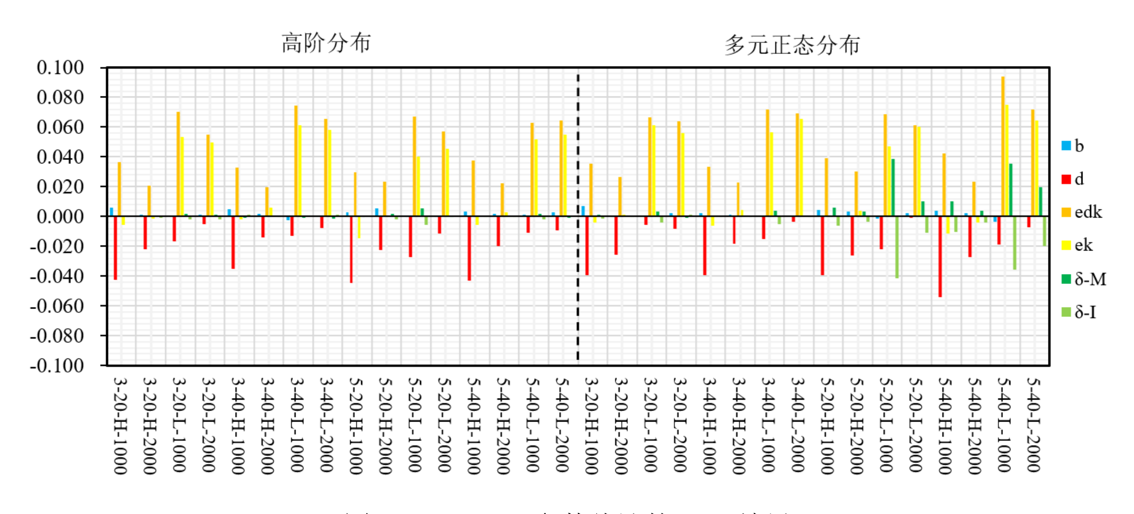


图 2 SDT-CDM 参数估计的 Bias 结果

注：*b* 为所有合理性参数的均值，*d* 为区分度参数，e_{DK} 为被试不会作答时的易度参数，ek 为被试会作答时的易度参数，δ-M 为属性的主效应，δ-I 为属性的交互效应。横坐标“3-20-H-1000 表示”3 属性-20 题-高题目质量-1000 人的实验条件。

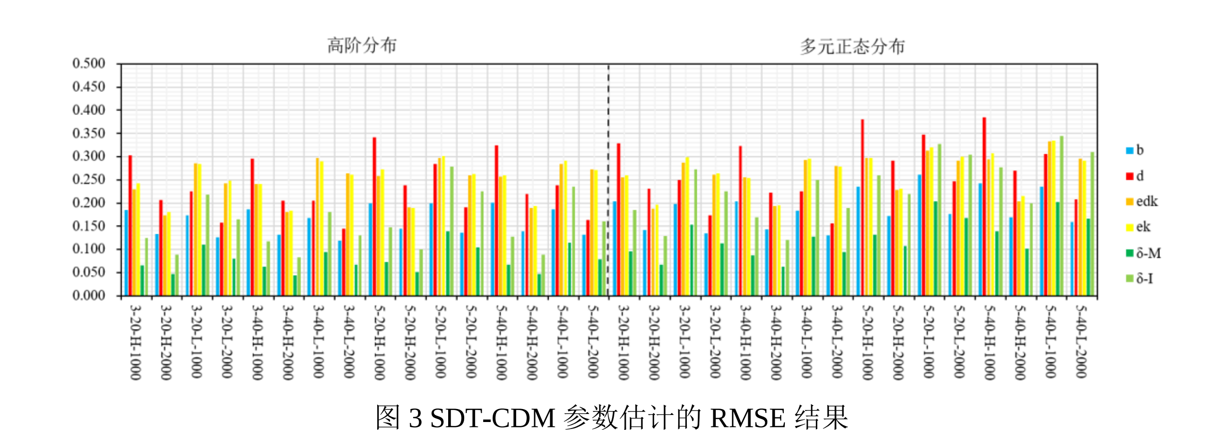


图 3 SDT-CDM 参数估计的 RMSE 结果

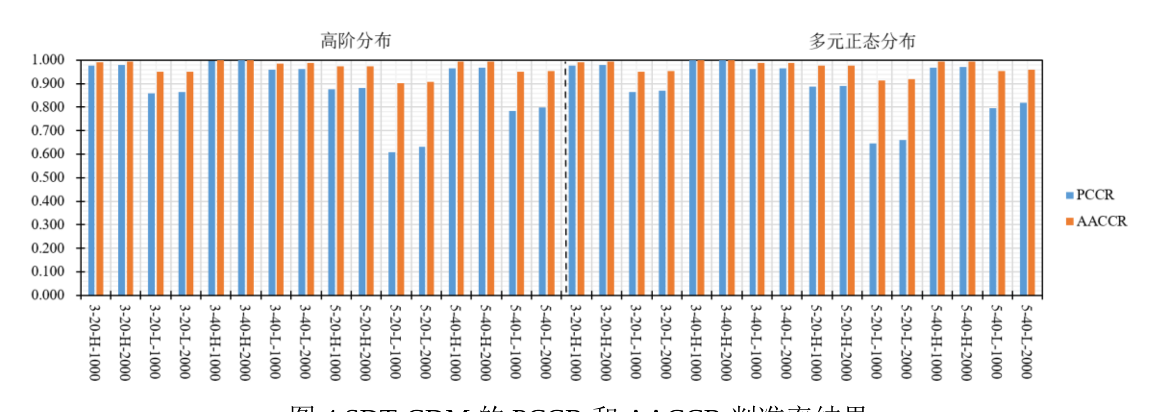


图 4 SDT-CDM 的 PCCR 和 AACCR 判准率结果

5 模拟研究 2

5.1 研究目的

采用蒙特卡洛模拟方式主要比较 SDT-CDM 和 NRDM 在不同实验条件下的被试分类准确性。NRDM 模型如下所示：

$$P_{jm}(Y_j = m_j|\boldsymbol{\alpha}_l) = \frac{\exp\left[\gamma_{0,j,m_j} + \boldsymbol{\gamma}_{j,m_j}^T \boldsymbol{h}(\boldsymbol{\alpha}_l, \boldsymbol{q}_j)\right]}{\sum_{m_j \in M_j} \exp\left[\gamma_{0,j,m_j} + \boldsymbol{\gamma}_{j,m_j}^T \boldsymbol{h}(\boldsymbol{\alpha}_l, \boldsymbol{q}_j)\right]} \quad (12)$$

其中，$\boldsymbol{\gamma}_{j,m_j}^T \boldsymbol{h}(\boldsymbol{\alpha}_l, \boldsymbol{q}_j) = \sum_{k=1}^{K} \gamma_{1,j,k,m_j}(\alpha_{l,k} q_{j,k}) + \sum_{k=1}^{K-1} \sum_{c=k+1}^{K} \gamma_{2,j,k,c,m_j}(\alpha_{l,k} \alpha_{l,c} q_{j,k} q_{j,c}) + \cdots$。$\gamma_{0,j,m_j}$为题目截距项，$\gamma_{1,j,k,m_j}$为题目的主效应，$\gamma_{2,j,k,c,m_j}$为题目的二阶交互，以此类推。关于 NRDM 的详细内容可参见 Templin 等(2008)文章。对比公式(2)和(12)不难看出，两个模型的题目参数含义不同且不能相互转化，如 SDT-CDM 为题目合理性参数b_{jm}等，NRDM 为题目主效应及各阶交互效应。此外，两个模型的参数范围也不同，SDT-CDM 的模型参数范围可以大于 1，而 NRDM 的模型参数范围均在 0-1 之间，这也使得两个模型的参数估计精度无法公平比较。因此，主要考察被试分类准确性的差异。

5.2 实验设计

本研究的自变量设置同研究 1。为了比较不同模型的表现差异，分别采用 SDT-CDM 和 NRDM 作为真模型生成数据，再用两个模型分别去拟合这些数据。其中，NRDM 的题目质量设置如下：高质量题目参数$1-P(\mathbf{1})$和$P(\mathbf{0})$从均匀分布$U[0.05, 0.15]$中随机抽取；低质量题目参数$1-P(\mathbf{1})$和$P(\mathbf{0})$从均匀分布$U[0.15, 0.25]$中随机抽取。$P(\mathbf{1})$和$P(\mathbf{0})$分别表示全部掌握和完全没掌握两种知识状态下的正确作答概率。其余设置同研究 1。

5.3 研究结果

网络版附录图 A1 和网络版附录图 A2 直观地呈现了两个模型分别为真模型时在不同自变量水平下的 PCCR 和 AACCR 结果。不论真模型是哪个，SDT-CDM 的表现均要优于 NRDM。当 SDT-CDM 为真模型时，属性分布对两个模型的分类精度影响均较小，样本量仅对 NRDM 有中等程度影响(样本量增大，N-PCCR 的均值提高了 7.6%)。属性个数由 3 个增加至 5 个时，S-PCCR 和 N-PCCR 的均值分别下降了 12.9%和 10.3%；题目质量降低时，S-PCCR 和 N-PCCR 的均值分别下降了 14.3%和 29.4%。值得注意的是，题目数量对两个模型的影响趋势存在不同，题目数量增大时，S-PCCR 的均值提高了 9.2%，但 N-PCCR 的均值反而下降了 18.2%。一个可能的原因是：题目数量越多，NRDM 的题目参数数量将大幅度增长(由公式(12)可以看出)，因此需要更多的样本量才能保证题目参数的估计精度，而当样本量不足时，题目参数的估计精度会降低，从而进一步降低了被试的分类精度。该影响也可以从最初提出 NRDM 的研究中得到佐证(Templin et al., 2008)，作者即使采用了缩减的补偿 NRDM 模型而非饱和

的 NRDM 模型也需要高达 5000 人才能得到理想的参数估计精度。相对而言，SDT-CDM 就表现的和大部分研究结果相近，即题目数量越多，分类精度越高，这点也可以说明新模型对于处理诊断测验中的称名数据更为理想。同时，这[6]也解释了为何 NRDM 作为真模型的表现仍不如 SDT-CDM。当 NRDM 为真模型时，尽管自变量对分类精度的影响趋势与真模型为 SDT-CDM 时类似，但此时 SDT-CDM 与 NRDM 的表现差异要更小，如题目质量降低时，S-PCCR 和 N-PCCR 的均值分别下降了 6.2%和 14.8%，这说明 SDT-CDM 比起 NRDM 具有更强的稳定性。

网络版附录表 A1 进一步呈现了不同自变量对两个模型差异的影响。不论真模型是哪个，题目数量对于两者的影响均是最大的，当 $J = 20$ 时，两者表现相差无几；但当 $J = 40$ 时，SDT-CDM 比 NRDM 的 PCCR 均值在不同真模型条件下分别高出了 42.29%和 21.04%，说明 NRDM 不太适合分析题目数量较多的测验，若要分析则需要增加较多样本量，而 SDT-CDM 在一定的样本量基础上就可以分析较多题量的测验情景。影响其次的是题目质量，尤其当题目质量较低时，SDT-CDM 比 NRDM 的 PCCR 均值在不同真模型条件下分别高出了 36.06%和 16.52%，说明 SDT-CDM 可以有效缓冲题目质量较低产生的负面影响。接下来是样本量，当样本量较小时，SDT-CDM 比 NRDM 的 PCCR 均值在不同真模型条件下分别高出了 24.72%和 14.93%，说明 SDT-CDM 比起 NRDM 来说更适合处理小样本。而其余变量均有不同程度的影响，不再赘述。

通过上述结果综合来看，SDT-CDM 从各方面都要优于 NRDM，通过详尽的模型比较研究，进一步证明了新模型的优势：当实验条件变化时，SDT-CDM 比 NRDM 更能维持住相对好的模型表现，因此可以认为 SDT-CDM 比 NRDM 的适用场景更广，表现更稳定。

[6] 即使是缩减的补偿 NRDM 模型也需要高达 5000 人才能得到理想的参数估计精度。

6 实证研究

实证数据取自 Ma 和 de la Torre(2020)使用过的 TIMSS 2011 数据，该数据共包含 23 道数学测验题目，本研究选择其中的 14 道选择题进行分析。数据中包含 748 名来自美国被试的作答数据，数据中的缺失值采用随机的错误答案进行替换。Q 矩阵属性个数为 6 个，分别为：A1)整数；A2)分数、小数和比例；A3)表达式、方程式和函数；A4)线条、角度和形状；A5)位置和移动；A6)数据组织、表示和解释识别明确信息，如表 1 所示。诊断结果的信效度指标采用 Wang 等(2015)提出的属性与模式分类一致性指标(Attribute-Level and Pattern-Level Classification Consistency)，以及属性与模式分类准确性(Attribute-Level and Pattern-Level Classification Accuracy)，它们可以分别从属性层面与模式层面综合判断诊断结果的信效度，均是取值越高则表明信效度越好。为了展现 SDT-CDM 的实际表现，在分析实证数据时加入了 NRDM[7]进行对比。

表 1 TIMSS 2011 数学测验(选择题)的 Q 矩阵

序号/题目编号	A1	A2	A3	A4	A5	A6
1/M032679	0	0	0	1	1	0
2/M042024	0	1	0	0	0	0
3/M042016	1	0	0	0	0	0
4/M042077	1	0	1	0	0	0
5/M042235	0	0	1	0	0	0
6/M042150	0	0	0	1	0	0
7/M032352	1	0	0	0	0	1
8/M032738	0	0	1	0	0	0
9/M032295	0	0	1	0	0	0
10/M032331	0	0	0	1	1	0
11/M042041	0	1	0	0	0	0
12/M032047	1	0	0	0	0	0
13/M032398	0	0	0	1	0	0
14/M032424	0	1	1	0	0	0

[7] 使用 R 软件中的 GDINA 程序包进行参数估计。

表2呈现了SDT-CDM与NRDM的模型-数据的相对拟合指标：负2倍对数似然值(-2 Log likelihood)、AIC(Akaike information criterion)与BIC(Bayesian information criterion)，三者均是取值越小越好。结果表明，SDT-CDM在3个拟合指标上的结果都要优于NRDM，如粗体结果所示，并且模型自由估计的参数数量为71个，而NRDM需要估计87个参数，更加复杂。

表2 模型数据相对拟合指标

Model	模型参数数量	-2LL	AIC	BIC
SDT-CDM	**71**	**19965.49**	**20107.49**	**20169.54**
NRDM	87	20007.68	20181.68	20257.71

网络版附录表A2和网络版附录表A3分别呈现了SDT-CDM和NRDM的模型参数估计结果。由网络版附录表A2可以看出，14道题目的区分度 d 均为正值，这表明“会答”题目的被试和“不会答”题目的被试能够被正常区分。理论上，d 越大则表明题目质量越好，但根据DeCarlo (2021)实证数据参数估计结果的经验，当 d 过大时可能导致标准误的增大，例如DeCarlo研究中 d 在6以上的3道题，其 d 值的标准误均在8以上，表明参数估计不稳定。相比之下，本研究仅有第7题的 d 值大于6，其标准误为4.044远小于8，整体来说，估计结果较为理想。

理论上，质量良好的题目的e_{DK}参数应该为负值且越小越好，对应到合理性参数b_1至b_4中的最大值可以是任意的干扰项，但不应该是正确选项，否则表示即使不会作答该题目的被试也能以较高的概率选中正确选项（即猜测概率高）。该测验中有9道题目的e_{DK}参数为负值，说明不会作答这些题目的被试感知到干扰项的合理性比正确选项的合理性更大，猜测行为发生的概率较小。但其余的5道题目的e_{DK}均为正值，说明正确选项比干扰项使得“不会答”的被试感觉到更合理，有较高的概率会发生猜测行为。以第 11 题为例，$e_{DK} = b_{正确选项} - b_{最大干扰项} = 2.227 - 1.391 = 0.836$，这说明对于“不会答”的被试而言，能以比较高的概率选择正确答案($\frac{e^{b_{jm}}}{\sum_{h=1}^{M} e^{b_{jh}}} = \frac{e^{2.227}}{e^{1.369}+e^{1.391}+e^{2.227}+e^{0}} = 0.509$)，因此可以认为第 11 题存在容易被猜对的问题，这与NRDM分析得到的结果(详细见表5所示)非常接近(对于没有掌握第11题所考察的属性的被试选择正确选项的概率较大，即P(0) = 0.498)，并且NRDM

对于其他题目猜测概率的估计与 SDT-CDM 模型也是高度一致的，这说明通过 SDT-CDM 的e_{DK}参数来判断题目的猜测行为是否过大是可行且准确的。

类似地，质量良好的题目的e_K应该为正且越大越好，对应到漂移过后的合理性参数b_1+dX_1至b_4+dX_4中的最大值需要是正确选项。再次以第 11 题为例，$e_K=\left(b_{正确选项}+d\right)-b_{最大干扰项}=e_{DK}+d=0.836+3.454=4.290$，表明该题目对于“会答”的被试能够以很高的概率选择正确选项($\frac{e^{b_{jm}+d_jX_{jm}}}{\sum_{h=1}^{M}e^{b_{jh}+d_jX_{jh}}}=\frac{e^{2.227+4.290}}{e^{1.369}+e^{1.391}+e^{2.227+4.290}+e^{0}}=0.987$)，符合逻辑。然而，第 12 题的$e_K$最小仅为 0.41，对于“会答”的被试来说，他们选择正确选项的概率仅为 0.358($=\frac{e^{0.765+0.410}}{e^{1.326}+e^{0.765+0.410}+e^{0.043}+e^{0}}$)，表明该题目存在较高的失误概率，需要对题目进行调整和完善。其余题目可以按照相同的方式进行分析后，用于判断题目/选项质量。

正如上述结果分析所示，SDT-CDM 可以指导测验编制者针对性地提高题目质量以及选项修改：通过e_{DK}的分析可知，这 14 道题中有 5 道题存在着猜测概率较大的问题，因此测验编制者需要编制更有诱导性/吸引力的干扰项。通过e_K的分析可知，所有的 11 道题均不存在逻辑异常的问题。但即使如此，测验编制者仍然可以根据 SDT-CDM 的分析结果针对性地对部分题目进行修改调整。例如举例的第 12 题，还有第 6 和 13 题的e_K均小于 1，并且这三道题的区分度d也是 14 道题中较低的，分别为 0.971，0.884 和 1.123。因此，若想进一步提高题目质量，可以尝试调整这两道题目的正确选项，增加“会答”与“不会答”时感知到的合理性差异(即d)。

网络版附录表 A4 是 SDT-CDM 的属性主效应和交互效应参数的估计结果。以第 1 题为例，$\delta_1=0.999$表明若被试仅掌握了题目 1 考察的第一个属性，其“会答”该题目的概率(即λ)为 99.9%，同理，若被试仅掌握了考察的第二个属性，其“会答”该题目的概率为 66.5%。而同时掌握了两个属性的被试，其“会答”该题目的概率相对于前两者分别提高了 0.1%与 33.5%。

表 3 呈现了属性与模式的分类准确性和分类一致性指标(Wang et al., 2015)结果。在分类准确性上，SDT-CDM 除 A1 属性低于 NRDM 之外，其余属性的分类准确性和模式分类准确性均要高于 NRDM，尤其是模式分类准确性提升了 39.13%，A6 的属性分类准确性提升了 23.77%；在分类一致性上，SDT-CDM 除 A1 属性低于 NRDM 之外，模式和其余属性的分类一致性均要高于 NRDM，尤其是 A6 的属性分类一致性提升了 28.63%。由表 2 的 Q 矩阵可知，A6 仅被考察了 1 次，相对其他属性被考察的次数偏少，此时对 NRDM 的影响更大，而

SDT-CDM 能够在有限考察次数内保持较高的分类准确性和一致性，更加稳健。以上结果表明新模型可以得到比旧模型更佳的信效度结果。

表 3 属性与模式水平的分类准确性和一致性

评价指标	模型	模式	属性					
			A1	A2	A3	A4	A5	A6
分类准确性	MC-SDT	**0.608**	0.864	**0.918**	**0.932**	**0.884**	**0.819**	**0.953**
	NRDM	0.437	**0.895**	0.907	0.930	0.875	0.780	0.770
	提升率	**39.13%**	-3.46%	**1.21%**	**0.22%**	**1.03%**	**5.00%**	**23.77%**
分类一致性	MC-SDT	**0.650**	0.809	**0.880**	**0.901**	**0.833**	**0.757**	**0.921**
	NRDM	0.647	0.850	0.866	0.895	0.823	0.720	0.716
	提升率	**0.46%**	-4.82%	**1.62%**	**0.67%**	**1.22%**	**5.14%**	**28.63%**

注：提升率 = (SDT-CDM - NRDM)/ NRDM

由于 SDT-CDM 能够报告传统 CDM 不能报告的难度和区分度参数，为了检查新模型所提供的难度与区分度参数的合理性，文章报告了其与两参数(2PL)和三参数(3PL)项目反应模型[8]的相关系数。由于 SDT-CDM 估计的是易度，在参数含义上与 2PL 和 3PL 中的难度(记作β)相反，因此需要将其反向。结果为：$r(-e_{DK}, \beta_{2PL}) = 0.63^{*}$，$r(-e_{DK}, \beta_{3PL}) = 0.71^{**}$，$r(-e_{K}, \beta_{2PL}) = 0.89^{***}$，$r(-e_{K}, \beta_{3PL}) = 0.79^{***}$。根据 Cohen(1988；P82)提出的标准，相关系数$r \geq 0.5$即为大效应量；此外根据张厚粲和徐建平(2015；P150)提出，相关系数在 0.6 至 0.8 之间即为强相关，0.8 以上即为非常强相关，并且以上 4 个相关系数均显著，因此表明新模型与 IRT 模型一样，都可以对题目进行难度表征，以此来反映题目的难度水平。由于 NRDM 无法表达难度参数，在 R 软件的 GDINA 程序包中可以提供广义区分度指标(global discrimination index, GDI; Xu et al., 2003)，因此仅报告 NRDM 与其他模型的区分度的相关结果：$r(d, a_{2PL}) = 0.66^{**}$，$r(d, a_{3PL}) = 0.79^{***}$，$r(\mathrm{GDI}, a_{2PL}) = 0.20^{ns}$，$r(\mathrm{GDI}, a_{3PL}) = 0.15^{ns}$，以上结果表明新模型估计得到的区分度参数 d 与 IRT 模型的估计结果为强相关，且均显著，但 NRDM 的区分度参数与 IRT 模型的结果相关较低且均不显著。

SDT-CDM 从可能的 64 种知识状态中识别出 748 名被试各自所属的知识状态。图 5 呈现了被试数量最多的前 10 类知识状态，总占比为 79.3%。进一步计算 SDT-CDM 和 NRDM 估计得到的属性掌握程度与总分间的相关(郭磊 等, 2021)，相关高表明总分越高

[8] 使用 R 软件的 MIRT 程序包进行参数估计。

的被试其掌握属性的程度越好，符合现实情况。其中，SDT-CDM 为 0.87***，NRDM 为 0.76***，表明新模型的表现要优于 NRDM。

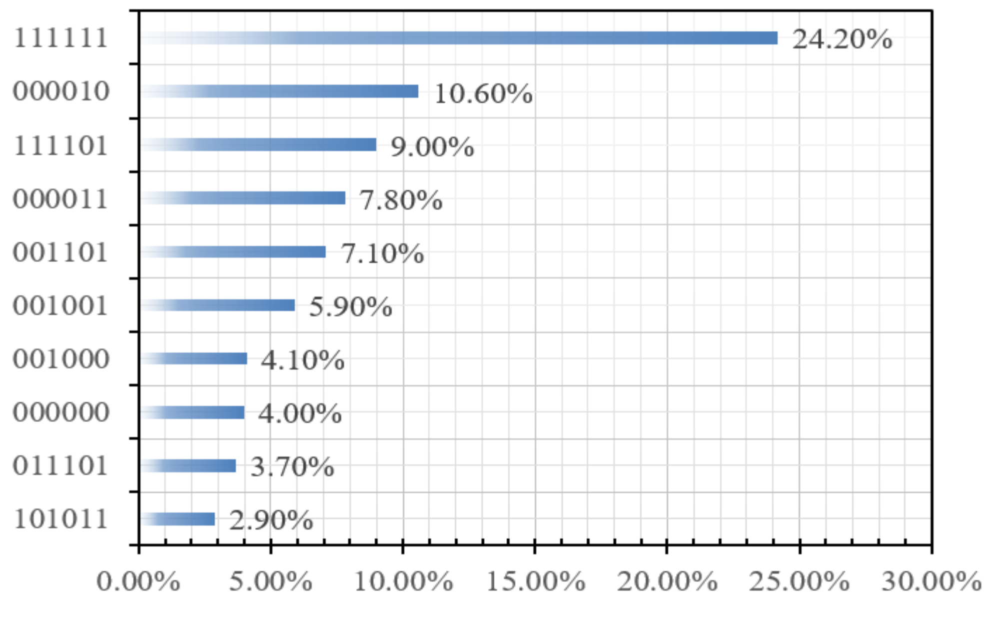


图 5 各类知识状态的被试占总体的比例(前十类)

6 讨论与研究结论

6.1 讨论与展望

MC 作答过程可以看作是信号检测的过程，意味着被试对每个选项都有一个合理性感知，并且总会选择感知到合理性最强的选项。本研究将 SDT 模型整合进 CDM 中，得到一些主要发现：首先，SDT-CDM 无需对 MC 题目的选项进行编码，而是为每个选项赋予了一个合理性参数，用来刻画选项之间的差异，并且通过这些合理性参数的组合可以计算得到传统诊断模型无法提供的难度和区分度参数，这些信息可用于题目质量诊断及修订。通过研究表明，SDT-CDM 的这些优势都是存在的，其模型构建是成功的。其次，通过两个模拟研究，在 5 个因素上全面地检验了新模型的性能，结果发现：(1)题目质量和样本量对 SDT-CDM 的参数估计精度影响较大，而属性分布、属性个数和题目数量的影响较小。(2)题目质量、属性个数和题目数量对被试判准率的影响较大，而属性分布和样本量对判准率的影响不大。(3)通过模型比较研究后发现，不论真模型是哪个，SDT-CDM 的被试判准率均要优于 NRDM，导致该现象的原因是由于 NRDM 需要很大样本量才能估计准确导致，这也恰恰证明了 SDT-CDM 的现实适用性和稳健性。最后，通过 TIMSS 2011 的实证数据分析发现，不论是模型数据拟合，分类精确性和一致性，还是与 IRT 的难度和区分度的相关，均是 SDT-CDM 表现更优。此外，由表 4 所得结果可用于判断题目/选项的质量和合理性，为完善和提升题

目质量提供的针对性指标，这也是 NRDM 所不能实现的功能。本研究值得探讨的问题还有以下几点。

6.1.1 干扰项信息的利用

目前国内外对于 MC 题型的认知诊断研究大部分都对干扰项进行了编码(de la Torre, 2009; DiBello et al., 2015; 郭磊 等, 2021; Ozaki, 2015; Wang et al., 2023)，这样可以充分利用干扰项所提供的诊断信息(即 q 向量信息)，将题目的诊断优势最大化。但正是该做法要求对选项层面进行编码，增大了题目编制的难度，此外若选项之间的 q 向量编码差异不大，亦或某些选项无法编码，其提供的额外诊断信息就变得有限。尽管 SDT-CDM 无需对干扰项进行编码，但新模型已将传统的 0-1 计分形式变为称名数据处理，而且提供了选项层面的参数(即合理性参数b_{jm})进行刻画，本质上这已经属于对选项层面信息的处理，并且通过模拟和实证研究表明，新模型的诊断分类准确性和一致性，以及模型拟合等结果均要优于 NRDM。本文可视作将 SDT 初次引进 CDA 领域的研究，未来可对 SDT-CDM 进行拓展，探索能将干扰项信息融入的新方法。一种潜在可行的思路是将混合参数λ_{lj}细化至选项层面，进一步刻画不同知识状态的被试与不同选项 q 向量之间的交互作用，以此综合反映被试“会作答”的可能性。

6.1.2 EM 算法的改进及标准误的计算

本研究推导了 SDT-CDM 的 EM 算法，但 EM 算法存在多样的变式(Chalmers, 2012)，例如标准的 EM 算法(the standard EM algorithm with fixed quadrature)、蒙特卡洛 EM 估计(Monte Carlo EM estimation)、随机 EM 算法(the stochastic EM)、MH-RM 算法(Metropolis-Hastings Robbins-Monro algorithm)、最小化卡方的 EM(朱玮，2006)等，这些算法大部分已应用于 IRT 研究领域，且可以通过 *mirt* 软件包实现。然而，目前在 CDM 中的 EM 算法比较单一，从 de la Torre(2009)提出 DINA 模型的边际极大似然的 EM 算法(MMLE/EM)后，MMLE/EM 便一直是主要的估计算法，包括本文也是使用这一框架拓展。尽管 MMLE/EM 算法简单高效，但探索精度更高、收敛更快、或具有其他独特优势的新算法很有必要。未来可以考虑将 IRT 里较为成熟的算法引入新模型中。

此外，CDM 中参数估计的标准误采用信息矩阵的逆求解，但目前已有多种信息矩阵(刘彦楼, 2022)，例如经验交叉相乘信息矩阵法(Empirical Cross-product Information Matrix, XPD)、观察信息矩阵法(Observed Information Matrix, Obs)和三明治信息矩阵法(Sandwich-type

Information Matrix, Sw)等。本文使用的是 XPD 矩阵，未来可探索使用不同信息矩阵对 SDT-CDM 参数标准误估计的影响。

6.1.3 与过程性数据相结合

随着计算机技术的发展，记录被试的作答过程性数据变得方便快捷，许多研究者开始挖掘这些过程性数据所提供的信息是如何帮助提升被试知识状态的诊断精度，以及反映出不同的作答风格或策略。如，和反应时数据结合的诊断(郑天鹏 等, 2023)，和眼动数据结合的诊断(詹沛达, 2022)，以及和动作序列结合的诊断(Zhan & Qiao, 2022)。这些研究均将过程性数据融入 CDM 中，并证明了融入辅助信息的可行性和有效性，为多模态数据分析提供了方法。尽管挖掘过程性数据中蕴含的信息已被研究者接受，但尚未就如何能更好地分析它们达成共识(He, et al., 2021)，同时，用于分析过程性数据本身的模型或方法也具有多样性，如处理计数数据的模型包括泊松模型(poisson model)、负二项式模型(negative binomial model)、零膨胀模型(zero-inflated model)、跨栏模型(Hurdle model)等。再如，动作序列的提取方法也有很多，如潜在空间模型(latent space model, Chen et al., 2022)，基于递归神经网的序列到序列自动编码器(recurrent neural network-based sequence-to-sequence autoencoders, Tang et al., 2021)，及多维尺度法(multidimensional scaling, Tang et al., 2020)等，不同的特征提取方法也会影响诊断分类的效果。未来可以探讨不同的过程性数据模型和不同的特征提取方法与 SDT-CDM 结合的实际效果。

6.1.4 与追踪诊断相结合

纵向追踪诊断研究也是 CDA 领域近年来的一个研究热点，通过对学习过程的追踪，不仅能进一步刻画学生的学习轨迹，更能有效发挥 CDA 的诊断功能，帮助教师等实施针对性补救教学，最终促进学生发展。目前纵向 CDM 包括基于潜在转移分析的纵向 CDM(Wang et al., 2018; Zhang & Chang, 2020)和基于高阶潜在结构的纵向 CDM 两大类(Lee, 2017; Zhan et al., 2019)，未来可以考虑将 SDT 模型融入纵向 CDM 中，不仅实现对被试知识状态的追踪，还能随时间点观察题目质量的改变。

本研究尚存一些不足之处，例如本研究只将 SDT-CDM 与 NRDM 进行比较，虽然这是由于能够处理选项层面数据且不需要选项层面编码的 CDM 较少导致，但正是缺乏更多的对比目标导致难以对 SDT-CDM 模型进行更深一步的探索研究。本文使用的 XPD 信息矩阵属于解析法信息矩阵，而解析法信息矩阵在计算 CDM 模型参数的标准误时可能会遇到矩阵非正定、以及方差协方差矩阵对角线元素可能小于 0 等问题，导致无法求解出标准误。因此计

算标准误更好的方法是采用刘彦楼(2022)提出的“并行自助法”，以类似于蒙特卡洛模拟的方式进行计算，可以不受解析法信息矩阵的限制，但本研究并未探索该方法在 SDT-CDM 模型中的有效性。此外，本文使用的 MMLE/EM 算法尽管高效，但 EM 算法可能会陷入局部最优解，Zeng 等(2023)提出了 Tensor-EM 算法，较好地改善了局部最优解的困境，对于复杂模型而言是很好的参数估计方法。

6.2 研究结论

本研究提出了基于信号检测论的认知诊断模型 SDT-CDM，基于模拟和实证研究结果，得出如下结论：

(1)SDT-CDM 可以通过 EM 算法实现其参数估计。除能提供传统诊断模型不能提供的题目难度和区分度参数外，还能估计得到每个选项的合理性参数，通过这些题目参数信息可以对题目进行修订以提高其质量。

(2)模拟研究结果表明，SDT-CDM 参数估计精度较好，不同自变量对题目参数和被试分类精度存在影响。其中，对分类精度影响重要性排序为：题目质量、属性个数和题目数量，而属性分布和样本量对精度的影响较小。

(3)实证研究结果表明，SDT-CDM 比 NRDM 有更好的模型数据拟合结果，更高的模式/属性分类准确性和一致性(尤其当某个属性被考察次数较少时，SDT-CDM 展现出了极高的稳定性)，被试属性总体掌握程度与其总分的相关结果也更高，且无需对干扰项进行编码。此外，可以根据两个易度参数(e_{DK}和e_K)和区分度参数 *d* 对题目质量进行诊断及针对性修订。

参考文献

Bock, R. D. (1972). Estimating item parameters and latent ability when responses are scored in two or more nominal categories. *Psychometrika, 37*, 29–51.

Bradshaw, L., & Templin, J. (2014). Combining item response theory and diagnostic classification models: a psychometric model for scaling ability and diagnosing misconceptions. *Psychometrika, 79*(3), 403–425.

Chalmers, R., P. (2012). mirt: A multidimensional item response theory package for the R environment. *Journal of Statistical Software, 48*(6), 1–29.

Chiu, C.-Y. (2013). Statistical refinement of the Q-matrix in cognitive diagnosis. *Applied Psychological Measurement, 37*(8), 598–618.

Chen, Y., Zhang, J., Yang, Y., & Lee, Y-S. (2022). Latent space model for process data. *Journal of Educational Measurement, 59*(4), 517–535.

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences (2nd ed.)*. New York, NY: Erlbaum.

DiBello, L. V., Henson, R. A., & Stout, W. F. (2015). A family of generalized diagnostic classification models for multiple choice option-based scoring. *Applied Psychological Measurement, 39*(1), 62–79.

DeCarlo, L, T. (2021). A signal detection model for multiple-choice exams. *Applied Psychological Measurement, 45*(6), 423–440.

de la Torre, J. (2009). DINA model and parameter estimation: a didactic. *Journal of Educational and Behavioral Statistics, 34*(1), 115–130.

de la Torre, J. (2011). The generalized DINA model framework. *Psychometrika*, *76*(2), 179–199.

Fang, G., Liu, J., & Ying, Z. (2019). On the identifiability of diagnostic classification models. *Psychometrika, 84*, 19–40.

Guo, L., Yuan, C. Y., & Bian, Y. F. (2013). Discussing the development tendency of cognitive diagnosis from the perspective of new models. *Advances in Psychological Science*, *21*(12), 2256–2264.

[郭磊，苑春永，边玉芳. (2013). 从新模型视角探讨认知诊断的发展趋势. *心理科学进展*, *21*(12), 2256–2264.]

Guo, L., Zheng C., Bian Y., Song N., & Xia L. (2016). New item selection methods in cognitive diagnostic computerized adaptive testing: combining item discrimination indices. *Acta Psychologica Sinica, 48*(7), 903–914.

[郭磊, 郑蝉金, 边玉芳, 宋乃庆, 夏凌翔. (2016). 认知诊断计算机化自适应测验中新的选题策略：结合项目区分度指标. *心理学报, 48*(7), 903–914.]

Guo, L., & Zhou, W. J. (2021). Nonparametric methods for cognitive diagnosis to multiple-choice test items. *Acta Psychologica Sinica, 53*(9), 1032–1043.

[郭磊, 周文杰. (2021). 基于选项层面的认知诊断非参数方法. *心理学报, 53*(9), 1032–1043.]

He, Q., Borgonovi, F., & Paccagnella, M. (2021). Leveraging process data to assess adults' problem-solving skills: using sequence mining to identify behavioral patterns across digital tasks. *Computers & Education, 166*: 104170.

Kelly, F. J. (1916). The kansas silent reading tests. *Journal of Educational Psychology, 7*, 63–80.

Lee, S. Y. (2017). *Growth curve cognitive diagnosis models for longitudinal assessment*. Unpublished doctoral thesis, University of California, Berkeley.

Liu, Y. (2022). Standard errors and confidence intervals for cognitive diagnostic models: parallel bootstrap methods. *Acta Psychologica Sinica, 54*(6), 703–724.

[刘彦楼. (2022). 认知诊断模型的标准误与置信区间估计：并行自助法. *心理学报*, *54*(6), 703–724.]

Ma, W., & de la Torre, J. (2016). A sequential cognitive diagnosis model for polytomous responses. *British Journal of Mathematical and Statistical Psychology, 69*(3), 253–275.

Ma, W., & de la Torre, J. (2020). An empirical Q-matrix validation method for the sequential generalized DINA model. *British Journal of Mathematical and Statistical Psychology, 73*(1), 142–163.

Ozaki, K. (2015). DINA models for multiple-choice items with few parameters: considering incorrect answers. *Applied Psychological Measurement*, *39*(6), 431–447.

Tang, X., Wang, Z., He, Q., Liu, J., & Ying, Z. (2020). Latent feature extraction for process data via multidimensional scaling. *Psychometrika, 85*, 378–397.

Tang, X., Wang, Z., Liu, J., & Ying, Z. (2021). An exploratory analysis of the latent structure of process data via action sequence autoencoder. *British Journal of Mathematical and Statistical Psychology, 74*(1), 1–33.

Templin, J., Henson, R., Rupp, A., Jang, E., & Ahmed, M. (2008). *Cognitive diagnosis models for nominal response data*. Annual Meeting of the National Council on Measurement in Education, New Brunswick, New Jersey.

Thissen, D., & Steinberg, L. (1997). *A response model for multiple-choice items*. In W. J. van der Linden & R. K. Hambleton (Eds.), Handbook of modern item response theory (pp. 51–65). Springer.

Wang, W., Song, L., Chen, P., Meng, Y., & Ding, S. (2015). Attribute-level and pattern-level classification consistency and accuracy indices for cognitive diagnostic assessment. *Journal of Educational Measurement, 52*, 457–476.

Wang, S., Yang, Y., Culpepper, S. A., & Douglas, J. A. (2018). Tracking skill acquisition with cognitive diagnosis models: a higher-order, hidden markov model with covariates. *Journal of Educational and Behavioral Statistics, 43*, 57–87.

Wang, Y., Chiu, C.-Y., & Kohn, H. F. (2023). Nonparametric classification method for multiple-choice items in cognitive diagnosis. *Journal of Educational and Behavioral Statistics, 48*(2), 189–219.

Xu, G. (2017). Identifiability of restricted latent class models with binary responses. *The Annals of Statistics, 45*, 675–707.

Xu, X., Chang, H., & Douglas, J. (2003). *A simulation study to compare CAT strategies for cognitive diagnosis*. Paper presented at the annual meeting of National Council on Measurement in Education, Montreal, Quebec, Canada.

Zeng, Z., Gu, Y. & Xu, G. (2023). A tensor-EM method for large-scale latent class analysis with binary responses. *Psychometrika, 88*, 580–612.

Zheng, T. P., Zhou, W. J., & Guo, L. (2023). Cognitive diagnosis modelling based on response times. *Journal of Psychological Science, 46*(2), 478–490.

[郑天鹏, 周文杰, 郭磊. (2023). 基于题目作答时间信息的认知诊断模型. *心理科学*, *46*(2), 478–490.]

Zhan, P. D. (2022). Joint-cross-loading multimodal cognitive diagnostic modeling incorporating visual fixation counts. *Acta Psychologica Sinica, 54*(11), 1416–1432.

[詹沛达. (2022). 引入眼动注视点的联合-交叉负载多模态认知诊断建模. *心理学报*, *54*(11), 1416–1432.]

Zhan, P. D., Jiao, H., Liao D. D., & Li, F. M. (2019). A longitudinal higher-order diagnostic classification model. *Journal of Educational and Behavioral Statistics*, *44*(3), 251–281.

Zhan, P. D., & Qiao, X. (2022). Diagnostic classification analysis of problem-solving competence using process data: an item expansion method. *Psychometrika, 87*(4), 1529–1547.

Zhang, H. C., & Xu, J. P. (2015). *Modern psychology and educational statistics (4th ed.)*. Beijing: Beijing Normal University Press.

[张厚粲, 徐建平. (2015). *现代心理与教育统计学(第4版)*. 北京：北京师范大学出版社.]

Zhang, S. S., & Chang, H. H. (2020). A multilevel logistic hidden markov model for learning under cognitive diagnosis. *Behavior Research Methods, 52*, 408–421.

Zhu W., Ding S., Chen X. (2006). Minimum chi-square/EM estimation under IRT. *Acta Psychologica Sinica, 38*(3), 453–460.

[朱玮, 丁树良, 陈小攀. (2006). IRT 中最小化 χ^2/EM 参数估计方法. *心理学报*, *38*(3), 453–460.]

Cognitive Diagnostic Assessment Based on Signal Detection Theory: Modeling and Application

Guo Lei1[1, 2]; QIN Haijiang[1, 3]

([1]Faculty of Psychology, Southwest University, Chongqing, China)

([2]Southwest University Branch, Collaborative Innovation Center of Assessment toward Basic Education Quality, Chongqing, China)

([3]Guiyang No.37 Middle School, Guiyang, China)

Abstract

Cognitive diagnostic assessment (CDA) is aimed at diagnose which skills or attributes examinees have or do not have as the name expressed. This technique provides more useful feedback to examinees than a simple overall score got from classical test theory or item response theory. In CDA, multiple-choice (MC) is one of popular item types, which have the superiority on high test reliability, being easy to review, and scoring quickly and objectively. Traditionally, several cognitive diagnostic models (CDMs) have been developed to analyze the MC data by including the potential diagnostic information contained in the distractors.

However, the response to MC items can be viewed as the process of extracting signals (correct options) from noises (distractors). Examinees are supposed to have perceptions of the plausibility of each options, and they make the decision based on the most plausible option. Meanwhile, there are two different states when examinee response to items: knows or does not know each item. Thus, the signal detection theory can be integrated into CDM to deal with MC data in CDA. The cognitive diagnostic model based on signal detection theory (SDT-CDM) is proposed in this paper and has several advantages over traditional CDMs. Firstly, it does not require the coding of *q*-vector for each option. Secondly, it provides discrimination and difficulty parameters that traditional CDMs cannot provide. Thirdly, it can directly express the relative differences between each options by plausibility parameters, providing a more comprehensive characterization of item quality.

The results of two simulation studies showed that (1) the marginal maximum likelihood estimation approach via Expectation Maximization (MMLE/EM) algorithm could effectively estimate the model parameters of the SDT-CDM. (2) the SDT-CDM had high classification

accuracy and parameter estimation precision, and could provide option-level information for item quality diagnosis. (3) independent variables such as the number of attributes, item quality, and sample size affected the performance of the SDT-CDM, but the overall results were promising. (4) compared with the nominal response diagnostic model (NRDM), the SDT-CDM was more accurate in classifying examinees under all data conditions.

Further, an empirical study on the TIMSS 2011 mathematics assessment were conducted using both the SDT-CDM and the NRDM to inspect the ecological validity for the new model. The results showed that the SDT-CDM had better fitting and a smaller number of model parameters than the NRDM. The difficulty parameters of the SDT-CDM were significantly correlated with those of the two- (three-) parameter logical models. And the same was true of the discrimination parameters for the SDT-CDM. However, the correlation between the discrimination parameters of the NRDM and those of the two- (three-) parameter logical models was low and not significant. Besides, the classification accuracy and classification consistency of the SDT-CDM were higher than those of the NRDM. All the results indicated that the SDT-CDM was worth promoting.

Keywords: signal detection theory, cognitive diagnostic assessment, multiple-choice items, expectation maximization algorithm